High-Order Modeling Techniques for Continuous Speech Recognition

Progress Report: 1 January 1995 – 31 March 1995

submitted to
Office of Naval Research
and
Advanced Research Projects Administration

by
Boston University
Boston, Massachusetts 02215

Principal Investigator

Dr. Mari Ostendorf
Associate Professor of ECS Engineering, Boston University
Telephone: (617) 353-5430

Administrative Contact

Maureen Rodgers, Awards Manager
Office of Sponsored Programs
Telephone: (617) 353-4365

Executive Summary

This research aims to develop new and more accurate stochastic models for speaker-independent continuous speech recognition by developing acoustic and language models aimed at representing high-order statistical dependencies within and across utterances, including speaker, channel and topic characteristics. These techniques, which have high computational costs because of the large search space associated with higher order models, are made feasible through a multi-pass search strategy that involves rescoring a constrained space given by an HMM decoding. With these overall project goals, the primary research efforts and results over the last quarter have included:

- explored sparse data training issues in dependence tree topology design;
- developed a parametric segment trajectory clustering algorithm;
- implemented a baseline HTK recognition system for a task of recognizing the Macrophone natural numbers data, achieving 76-78% accuracy; and
- began experiments assessing two auditory signal processing models in phone recognition on the TIMIT corpus.

We also re-evaluated our November 1994 ARPA benchmark recognition system using a new N-best list provided by BBN, and achieved 10% word error rate (vs. 11.6% in the official benchmark tests).

Note that in this reporting period, the project efforts have been temporarily scaled back while Prof. Ostendorf is on sabbatical in Japan and supervising the research remotely.

Contents

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: High-Order Modeling Techniques for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1995 – 31 March 1995

1 Productivity Measures

- Refereed papers submitted but not yet published: 0
- Refereed papers published: 0
- Unrefereed reports and articles: 1
- Books or parts thereof submitted but not yet published: 0
- Books or parts thereof published: 0
- Patents filed but not yet granted: 0
- Patents granted (include software copyrights): 0
- Invited presentations: 0
- Contributed presentations: 0
- Honors received: none
- Prizes or awards received: none
- Promotions obtained: none
- Graduate students supported $\geq$ 25% of full time: 1
- Post-docs supported $\geq$ 25% of full time: 0
- Minorities supported: 1 woman

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: High-Order Modeling Techniques for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1995 – 31 March 1995

2 Summary of Technical Progress

Introduction and Background

The goal of this work is to develop and explore novel stochastic modeling techniques for acoustic and language modeling in large vocabulary continuous speech recognition, particularly recognition of spontaneous speech. To answer the challenges posed by spontaneous speech recognition requires improvements at all levels of the recognition process: signal processing, acoustic modeling, lexical representation and language modeling – both in terms of the baseline stochastic models and the techniques for adapting these models. In addressing these challenges, the general theme of our research is high-level correlation modeling, i.e. representing correlation of observations beyond the level of the frame or the word to dependencies within and across utterances associated with speaker, channel, topic and/or speaking style. In particular, we will concentrate on three problems: hierarchical intra-utterance dependence modeling, unsupervised adaptation of acoustic models within and across utterances, and sub-language modeling triggered by both acoustic and dialog-level cues.

This project builds on work done at BU funded by previous ARPA-NSF and ARPA-ONR grants on continuous speech recognition. Acoustic modeling is based on the stochastic segment model (SSM) [?, ?], where we hope to benefit from our previous efforts in distribution clustering [?], automatic distribution mapping, adaptation, and dynamical system modeling [?]. In addition, a corner stone of the high-order modeling work will be the use of dependence trees, exploiting developments in [?] which we hope will provide a mechanism for capturing speaker-dependent effects in a speaker-independent model. In language modeling, we build on our sentence-level mixture language model [?] and dynamic cache [?] work, as well as incorporate conditioning on prosody and dialogue cues. Finally, we rely on a multi-pass recognition search strategy [?, ?], which reduces the search space with standard hidden Markov models in order to allow rescoring with the higher-order (and therefore more computational) models proposed here. Thus, our work builds on existing strengths of speech recognition technology, while exploring radically new knowledge sources (i.e. long-term dependence), a combination that has the potential to significantly advance the state of the art in speech recognition performance.

Summary of Technical Results

The research efforts during this period, supported in part by AASERT awards from ONR and ARPA, have focused on issues in automatic training of the hierarchical model, further development of the theory of segmental modeling, and establishing a baseline telephone speech recognition result. These efforts and recent results on the November 1994 ARPA benchmark test are summarized below. Note that in this reporting period, the project efforts have been temporarily scaled back while Prof. Ostendorf is on sabbatical in Japan and supervising the research remotely.

Intra-utterance phoneme dependence modeling. Over the past year, we have developed the theoretical framework for a hierarchical model of dependence for a set of discrete random variables, which we plan to use as a model of intra-utterance phoneme dependence. We use a dependence tree [?] to represent the correlation among random variables, i.e. a tree structure (designed automatically) with Markov assumptions along the branches of the tree. The dependence tree can be thought of as representing a vector "state" that describes the speaker/utterance, where each element of the vector corresponds to a phoneme. Since most utterances will not contain all possible phonemes, we derived an efficient algorithm for computing the likelihood of the observed data [?], which we call the upward-downward algorithm to emphasize the analogy to the forward-backward algorithm. This algorithm is needed in the E-step in the EM parameter estimation algorithm, which is one step in our iterative approach to combined dependence tree topology design and parameter estimation. In previous work, we implemented the training algorithm for discrete distribution dependence trees (iterative topology and parameter estimation), and began initial experiments on the TIMIT corpus. In this quarter, we have continued experimentation, focusing on problems of robust estimation and comparison of automatically designed vs. hand-specified tree topologies. We discovered topology design problems due to a bias in the mutual information estimate for infrequently observed classes and are investigating various methods to compensate for differences in training set sizes for different pairs of phones. In parallel with the experimental work, we have begun looking at extension of the upward-downward algorithm to Gaussian and Gaussian mixture distributions. [This work was due to graduate student Orith Ronen.]

Segmental modeling theory. As part of Prof. Ostendorf's sabbatical, she has been working on refining an integrated theory of stochastic modeling for speech recognition, combining segment models and hidden Markov models. In this work, she extended the maximum likelihood distribution clustering approach to parameter tying [?] to the parametric trajectory segmental models of Gish and Ng [?] and is currently working with colleagues at ATR on the use of this algorithm for automatic discovery of segmental units for sub-word modeling. [This work was supported in part by ATR Interpreting Telecommunications Laboratories.]

Recognition of telephone speech. As mentioned in the previous report, we have decided to use the Macrophone Natural Number corpus [?] and an HTK HMM recognition system as a test paradigm for channel modeling research for telephone speech recognition. In this reporting period, baseline results for this task were obtained with standard signal processing techniques, using 13 cepstral features plus their derivatives and cepstral mean subtraction. The system used a word-pair grammar, a 530-word phonemic pronunciation dictionary, and 3-state, left-to-right HMMs with diagonal covariance Gaussian distributions. We used an incremental model building approach [?], first training monophones, then triphones with clustered states, and then incrementing the number of mixtures in the triphone distributions. Specifically, preliminary monophone models were first developed using 47 phones and two non-speech models (noise and begin/end of utterance silence). Then triphone models were trained using 405 state-clustered triphones and three non-speech models (noise, inter-word pause/breath/silence, and a begin/end of utterance silence). The best results obtained so far for both cases used five mixtures per state. On independent development test sets of Macrophone natural numbers divided into two even test sets, the monophone system achieved 57.4% and 58.9% word accuracy and the triphone system achieve 76.4% and 78.3% accuracy. We may continue some experimentation with the baseline system, but next plan to evaluate RASTA processing [?] and Bayesian channel estimation as an alternative to cepstral mean subtraction. [This work was supported by an ARPA AASERT award associated with this project and was due to undergraduate student John Kaufhold and graduate student Rebecca Bates.]

Auditory-based feature extraction. It has often been suggested that auditory-based signal processing might make a better front end for speech recognition than cepstral analysis, but few if any models have led to improved results on high quality speech recordings. One reason for this lack of encouraging results might be that earlier studies used Gaussian distributions that may not be well suited to auditory features. Thus, we decided to assess use of a standard auditory model, the Seneff model [?], with a (relatively) non-parametric distribution assumption, i.e. mixtures of Gaussians which are tied across states at the full distribution level and not the component Gaussian level. Experiments on a small task of recognizing broad classes of phonemes with context-independent models gives 58% accuracy for the Seneff model, compared to 65% for cepstra. Using a cosine transform plus derivatives on the auditory features gave only a small improvement. Full-scale phone recognition experiments on the TIMIT corpus are currently running, using the system building strategy outlined in [?]. A second reason why auditory models have not yielded bigger gains may be that there is other information in the auditory process that is not being taken advantage of. To assess this possibility, we are working with various stages of the Carney auditory model [?], currently modifying the outputs to provide a data rate appropriate for speech recognition. Eventually, we plan to compare all three processors on the TIMIT phone recognition task. [This work was supported by an ONR AASERT award associated with this project and was due to undergraduate student Susan Zlotkin.]

ARPA benchmark tests and language modeling. In the November 1994 ARPA benchmark tests, we reported a best case 10.9% word error rate (post-adjudication), but these results were based on rescoring N-best lists on a preliminary version of the BBN recognition system which was not the same as that which they used in the official benchmark. After the benchmark, we ran our system on the N-best lists from the BBN benchmark system and achieved a best case word error of 10.0%, which compares with 10.2% reported by BBN and shows a small gain from the additional use of the stochastic segment model. The initial results reflected a mismatch between acoustic model and language model lexicons, which was somewhat improved by using the new N-best lists. We had hoped that this mismatch would explain why our various language modeling advances did not yield improvements in recognition results, despite significant improvements in perplexity (15-30%). We did observe language modeling gains with the new N-best lists on the development test set, but still no improvement in performance on the evaluation test set. We hope that further experiments will lead us to a better understanding of these results; however, we observed that other sites working on language modeling also did not achieve significant gains in recognition performance.

Future Goals

The low level of effort on this project will continue into the next quarter, during which time we plan to concentrate on the hierarchical dependence tree model and implementation of the channel estimation algorithm. Over the summer, we plan to increase the level of effort on this project and focus on recognition of the spontaneous speech in the Switchboard Corpus.

References

[1] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. on Acoust., Speech, and Signal Proc.,* 1989.

[2] S. Roucos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.,* pp. 127–130, New York, New York, April 1988.

[3] A. Kannan, M. Ostendorf and J. R. Rohlicek, "Maximum Likelihood Clustering of Gaussians for Speech Recognition," *IEEE Trans. on Speech and Audio Proc.,* Vol. 2, No. 3, 1994, pp. 453-455.

[4] V. Digalakis, J. R. Rohlicek and M. Ostendorf, "A Dynamical System Approach to Continuous Speech Recognition," *IEEE Trans. on Speech and Audio Proc.,* Vol. 1, No. 4, 1993, pp. 431-442.

[5] O. Ronen, J. R. Rohlicek and M. Ostendorf, "Parameter Estimation of Dependence Tree Models Using the EM Algorithm," *IEEE Signal Processing Letters,* to appear July 1995.

[6] R. Iyer, M. Ostendorf and J. R. Rohlicek, "Language Modeling with Sentence-Level Mixtures," *Proc. ARPA Workshop on Human Language Technology,* March 1994.

[7] M. Ostendorf, F. Richardson, R. Iyer, A. Kannan, O. Ronen and R. Bates, "The 1994 BU NAB News Benchmark System," *Proceedings of the ARPA Workshop on Spoken Language Technology,* January 1995, pp. 139-142.

[8] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, J. R. Rohlicek, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses," *Proc. DARPA Workshop on Speech and Natural Language,* pp. 83-87, February 1991.

[9] F. Richardson, M. Ostendorf and J. R. Rohlicek, "Lattice-based Search Strategies for Large Vocabulary Speech Recognition," *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.,* 1995.

[10] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory,* Vol. IT-14, No. 3, May 1968, pp. 462-467.

[11] H. Gish and K. Ng, "A Segmental Speech Model with Applications to Word Spotting," in *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.,* 1993, pp. II-447-450.

[12] K. Taussig and J. Bernstein, "Macrophone: An American English Telephone Speech Corpus," *Proc. ARPA Spoken Language Technology Workshop,* 1994.

[13] H. Hermansky, N. Morgan and H. Hirsch, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.,* 1992, ppII83-86.

[14] S. Seneff, "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing", *Journal of Phonetics*, vol. 16, pp. 55–76, 1988.

[15] P.C. Woodland and S.J. Young, "The Use of State Tying in Continuous Speech Recognition", *Proc. European Conf. on Speech Communication and Technology,* 1993.

[16] L. H. Carney, "A Model for the Responses of Low-Frequency Auditory-Nerve Fibers in Cat," *J. Acoust. Soc. Am.,* vol. 93, pp. 401–417, 1993.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: High-Order Modeling Techniques for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1995 – 31 March 1995

3 Publications and Presentations

During this reporting period, we published one conference paper.

Unrefereed Reports and Conference Papers:

"The 1994 BU NAB News Benchmark System," M. Ostendorf, F. Richardson, R. Iyer, A. Kannan, O. Ronen and R. Bates, *Proceedings of the ARPA Workshop on Spoken Language Technology,* January 1995.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: High-Order Modeling Techniques for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1995 – 31 March 1995

4 Transitions and DoD Interactions

The initial grant included a subcontract to BBN, and the research results and software are available to them. Thus far, we have collaborated with BBN by combining the Byblos system with the SSM in N-Best hypothesis rescoring to obtain improved recognition performance, and we have provided BBN with papers and technical reports to facilitate sharing of algorithmic improvements. In addition, Boston University student Fred Richardson has implemented software libraries that will be shared by both sites, and he has modified the BBN decoder to provide lattices annotated with segmentation times and HMM scores. BU students Fred Richardson and Rukmini Iyer are currently working at BBN, while Prof. Ostendorf is on sabbatical in Japan, and part of their efforts are aimed at porting software from BU to BBN.

The recognition system that has been developed under the support of this grant and of a joint NSF-ARPA grant (NSF # IRI-8902124) is currently being used for automatically obtaining good quality phonetic alignments for a corpus of radio news speech under development at Boston University in a project supported by the Linguistic Data Consortium.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: High-Order Modeling Techniques for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 January 1995 – 31 March 1995

5 Software and Hardware Prototypes

Our research has required the development and refinement of software systems for parameter estimation and recognition search, which are implemented in C or C++ and run on Sun Sparc workstations. No commercialization is planned at this time.

OFFICE OF THE UNDER SECRETARY OF DEFENSE (ACQUISITION)
DEFENSE TECHNICAL INFORMATION CENTER
CAMERON STATION
ALEXANDRIA, VIRGINIA 22304-6145

IN REPLY REFER TO DTIC-OCC

SUBJECT: Distribution Statements on Technical Documents

TO: OFFICE OF NAVAL RESEARCH
CORPORATE PROGRAMS DIVISION
ONR 353
800 NORTH QUINCY STREET
ARLINGTON, VA 22217-5660

1995 1027 014

1. Reference: DoD Directive 5230.24, Distribution Statements on Technical Documents, 18 Mar 87.

2. The Defense Technical Information Center received the enclosed report (referenced below) which is not marked in accordance with the above reference.

QUARTERLY REPORT/1JAN-31 MAR95
N00014-92-J-1778
TITLE: HIGH-ORDER MODELING
TECHNIQUES FOR CONTINUOUS
SPEECH RECOGNITION

3. We request the appropriate distribution statement be assigned and the report returned to DTIC within 5 working days.

4. Approved distribution statements are listed on the reverse of this letter. If you have any questions regarding these statements, call DTIC's Cataloging Branch, (703) 274-6837.

FOR THE ADMINISTRATOR:

1 Encl

GOPALAKRISHNAN NAIR
Chief, Cataloging Branch

FL-171
Jul 93

DISTRIBUTION STATEMENT A:

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION IS UNLIMITED

DISTRIBUTION STATEMENT B:

DISTRIBUTION AUTHORIZED TO U.S. GOVERNMENT AGENCIES ONLY; (Indicate Reason and Date Below). OTHER REQUESTS FOR THIS DOCUMENT SHALL BE REFERRED TO (Indicate Controlling DoD Office Below).

DISTRIBUTION STATEMENT C:

DISTRIBUTION AUTHORIZED TO U.S. GOVERNMENT AGENCIES AND THEIR CONTRACTORS; (Indicate Reason and Date Below). OTHER REQUESTS FOR THIS DOCUMENT SHALL BE REFERRED TO (Indicate Controlling DoD Office Below).

DISTRIBUTION STATEMENT D:

DISTRIBUTION AUTHORIZED TO DOD AND U.S. DOD CONTRACTORS ONLY; (Indicate Reason and Date Below). OTHER REQUESTS SHALL BE REFERRED TO (Indicate Controlling DoD Office Below).

DISTRIBUTION STATEMENT E:

DISTRIBUTION AUTHORIZED TO DOD COMPONENTS ONLY; (Indicate Reason and Date Below). OTHER REQUESTS SHALL BE REFERRED TO (Indicate Controlling DoD Office Below).

DISTRIBUTION STATEMENT F:

FURTHER DISSEMINATION ONLY AS DIRECTED BY (Indicate Controlling DoD Office and Date Below) or HIGHER DOD AUTHORITY.

DISTRIBUTION STATEMENT X:

DISTRIBUTION AUTHORIZED TO U.S. GOVERNMENT AGENCIES AND PRIVATE INDIVIDUALS OR ENTERPRISES ELIGIBLE TO OBTAIN EXPORT-CONTROLLED TECHNICAL DATA IN ACCORDANCE WITH DOD DIRECTIVE 5230.25, WITHHOLDING OF UNCLASSIFIED TECHNICAL DATA FROM PUBLIC DISCLOSURE, 6 Nov 1984 (Indicate date of determination). CONTROLLING DOD OFFICE IS (Indicate Controlling DoD Office).

The cited documents has been reviewed by competent authority and the following distribution statement is hereby authorized.

A
(Statement)

OFFICE OF NAVAL RESEARCH
CORPORATE PROGRAMS DIVISION
ONR 353
800 NORTH QUINCY STREET
ARLINGTON, VA 22217-5660

(Controlling DoD Office Name)

(Reason)

(Controlling DoD Office Address, City, State, Zip)

DEBRA T. HUGHES
DEPUTY DIRECTOR
CORPORATE PROGRAMS OFFICE

25 SEP 1995

(Signature & Typed Name) (Assigning Office)

(Date Statement Assigned)